TOPIC 2

NUTRITION

Name: Akash Mandal

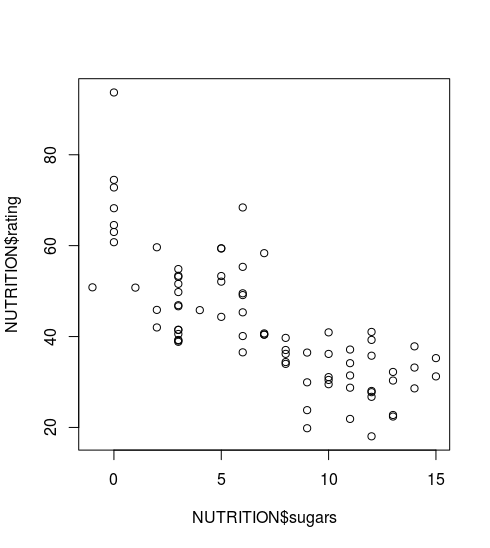
Roll No. : 16MA20007

Course: CS40003

Rating and Sugar

We may think of some correlation between rating and sugar content, and we draw a scatter plot to visualise the type of correlation.

> plot(NUTRITION$sugars, NUTRITION$rating)



Linear Model

We try to fit a linear model to the data, and hence calculate correlation, and R-squared values.

r = -0.7596747, R2 = 0.5771

This, of course denotes a moderately strong negative correlation.

We get the regression line as -

rating = 59.2844 + (-2.4008) \* sugars

> cor(NUTRITION$rating, NUTRITION$sugars)

[1] -0.7596747

> linearModel <- lm(formula = rating ~ sugars, data = NUTRITION)

> summary(linearModel)

Call:

lm(formula = rating ~ sugars, data = NUTRITION)

Residuals:

Min 1Q Median 3Q Max

-17.853 -5.677 -1.439 5.160 34.421

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 59.2844 1.9485 30.43 < 2e-16 \*\*\*

sugars -2.4008 0.2373 -10.12 1.15e-15 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.196 on 75 degrees of freedom

Multiple R-squared: 0.5771,Adjusted R-squared: 0.5715

F-statistic: 102.3 on 1 and 75 DF, p-value: 1.153e-15

Quadratic model

We try a quatratic fit (non linear regression) for the data, and assume sugar2 as a second variable in a multiple regression problem.

We get the regression relation as

rating = 63.33743 + (-4.23661) \* sugars + 0.13761 \* sugars2

> #quadratic fit

> sugars2 <- sugars^2

> twodegree <- lm(rating ~ sugars + sugars2)

> summary(twodegree)

Call:

lm(formula = rating ~ sugars + sugars2)

Residuals:

Min 1Q Median 3Q Max

-16.9733 -4.7005 -0.3318 4.5481 30.3675

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 63.33743 2.51298 25.204 < 2e-16 \*\*\*

sugars -4.32661 0.82134 -5.268 1.31e-06 \*\*\*

sugars2 0.13761 0.05635 2.442 0.017 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.906 on 74 degrees of freedom

Multiple R-squared: 0.6086,Adjusted R-squared: 0.5981

F-statistic: 57.54 on 2 and 74 DF, p-value: 8.413e-16

Cubic Model

Finally we try a cubic fit to our data, in a similar procedure as the quadratic fit. We obtain the following regression relation.

rating = 62.958876 + (-3.854666) \* sugars + 0.047327 \* sugars2 + 0.004325 \* sugars3

> #cubic fit

> sugars3 <- sugars^3

> threedegree <- lm(rating ~ sugars + sugars2 + sugars3)

> summary(threedegree)

Call:

lm(formula = rating ~ sugars + sugars2 + sugars3)

Residuals:

Min 1Q Median 3Q Max

-16.0282 -5.0420 -0.3263 4.7912 30.7460

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 62.958876 2.775380 22.685 <2e-16 \*\*\*

sugars -3.854666 1.649255 -2.337 0.0222 \*

sugars2 0.047327 0.278884 0.170 0.8657

sugars3 0.004325 0.013080 0.331 0.7419

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.96 on 73 degrees of freedom

Multiple R-squared: 0.6092,Adjusted R-squared: 0.5932

F-statistic: 37.94 on 3 and 73 DF, p-value: 6.903e-15

Plot

We now plot the three regression models on the scatterplot to visually examine how well our regression lines fit the given data.

Yellow: Linear

Blue: Quadratic

Red: Cubic

> xplot = seq(-1, 16, 0.1)

> predictedrating2 <- predict(twodegree, list(sugars=xplot, sugars2=xplot^2))

> predictedrating3 <- predict(threedegree, list(sugars=xplot, sugars2=xplot^2, sugars3=xplot^3))

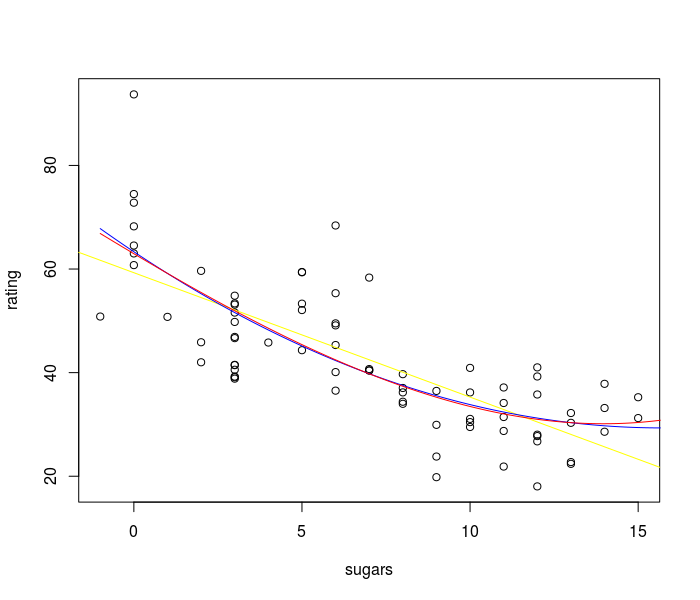
>

> plot(sugars, rating)

> abline(linearModel, col="yellow")

> lines(xplot, predictedrating2, col="blue")

> lines(xplot, predictedrating3, col="red")



Conclusion

We note the multiple and adjusted r-squared values computed by our models, in the summary statistics. We note that the quadratic model, with an adjusted r-squared of 0.5981, is the best probable fit to the data.

Quadratic regression explains around 60% of the variation in rating due to sugar.